Learning More from Mixed Emotions: A Label Refinement Method for Emotion Recognition in Conversations

Jintao Wen*, **Geng Tu***, **Rui Li**, **Dazhi Jiang**,† **Wenhua Zhu**†
Department of Computer Science
Shantou University, China
{20jtwen, 19gtu, ruili, dzjiang, 21whzhu}@stu.edu.cn

Abstract

One-hot labels are commonly employed as ground truth in Emotion Recognition in Conversations (ERC). However, this approach may not fully encompass all the emotions conveyed in a single utterance, leading to suboptimal performance. Regrettably, current ERC datasets lack comprehensive emotionally distributed labels. To address this issue, we propose the Emotion Label Refinement (EmoLR) method, which utilizes context- and speaker-sensitive information to infer mixed emotional labels. EmoLR comprises an Emotion Predictor (EP) module and a Label Refinement (LR) module. The EP module recognizes emotions and provides context/speaker states for the LR module. Subsequently, the LR module calculates the similarity between these states and ground-truth labels, generating a refined label distribution (RLD). The RLD captures a more comprehensive range of emotions than the original one-hot labels. These refined labels are then used for model training in place of the one-hot labels. Experimental results on three public conversational datasets demonstrate that our EmoLR achieves state-of-the-art performance.

1 Introduction

Emotion recognition in conversations (ERC) is an important research topic with broad applications, including human–computer interaction (Poria et al., 2017), opinion mining (Cambria et al., 2013), and intent recognition (Ma et al., 2018).

Unlike vanilla text emotion detection, ERC models need to model context- and speaker-sensitive dependencies (Tu et al., 2022a) to simulate the interactive nature of conversations. Recurrent neural networks (RNNs) (Zaremba et al., 2014) and their variants have been successfully applied for ERC. Recently, ERC research has primarily focused on understanding the influence of internal/external factors on emotions in conversations, such as topics (Zhu et al., 2021), commonsense (Zhong et al., 2019; Jiang et al., 2022), causal relations (Ghosal et al., 2020), and intent (Poria et al., 2019b). These efforts have improved the model's understanding of the semantic structure and meaning of conversations.

However, despite advancements in modeling context information (Zhong et al., 2021; Saxena et al., 2022), the final classification paradigm in ERC remains the same: calculating a certain loss between predicted probability distribution and one-hot labels. This black-and-white learning paradigm has the following problem: According to Plutchik's wheel of emotions (Plutchik, 1980) and the hourglass model (Cambria et al., 2012), emotions expressed in dialogue are often mixed expressions that include various basic emotions (such as anger, sadness, fear, and happiness) (Chaturvedi et al., 2019; Jiang et al., 2023). Each basic emotion contributes to the overall emotional expression to some extent. However, the one-hot label representation assumes emotions are independent. In real scenarios, as shown in Figure 1, the dialogue expression is normally not a single emotion, but multiple emotions are presented with a specific distribution. Particularly when dealing with mixed and ambiguous emotions, one-hot vectors as labels are insufficient. They will overlook the emotional information within the utterance, which may lead to suboptimal performance of the ERC models.

To address the limitations of one-hot labels, several efforts have been proposed in the fields of text classification and object detection, such as label smoothing (LS) (Müller et al., 2019) and label distribution learning (LDL) (Geng, 2016). LS partially alleviates the problem by randomly injecting noise, but it still cannot fundamentally recover the inherent emotional distribution within an

*Both authors contribute equally to this work.
†Corresponding authors.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1485–1499, 2023. https://doi.org/10.1162/tacl_a_00614
Action Editor: Yang Liu. Submission batch: 8/2022; Revision batch: 8/2023; Published 12/2023.

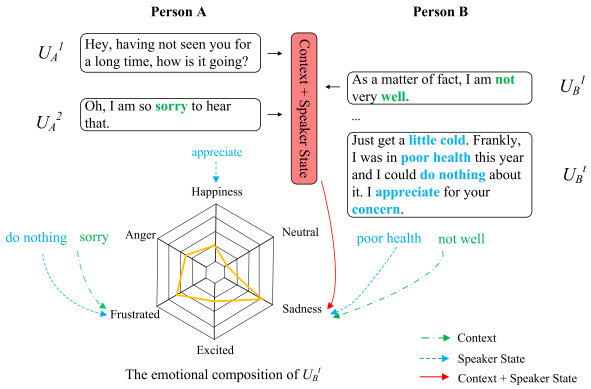


Figure 1: The emotions expressed in an utterance are abundant and interrelated. The left-bottom Radar chart shows the emotion of U_B^t utterance. The emotion is influenced by the context (green lines) and the current speaker's state (blue lines). The red lines indicate that the emotion is recognized by both context and speaker states. For example, U_B^t not only expresses sadness but also includes frustration and happiness. Although sad emotion has the dominant effect in this utterance, we cannot ignore the semantic information contained in other emotions.

utterance. LDL is capable of handling instances with multiple labels quantitatively, making it advantageous for tasks involving fuzzy labels (Xu et al., 2019). However, obtaining the true label distribution for structured conversation data is challenging.

Based on this, we propose the Emotion Label Refinement (EmoLR) method based on context- and speaker-sensitive information. EmoLR consists of two components: the emotion predictor (EP) and the label refinement (LR). The EP module preserves context- and speaker-related representations during the emotion detection process. In the LR module, the context- and speaker-sensitive representations are compared with each label to estimate their correlation individually. The refined label is generated based on the correlation scores. The original one-hot label is combined with the new label distribution to reduce interference from noise in the model. The final output, after softmax activation, is used for training. The refined label, capturing the relationship between the speaker and contextual information, reflects all possible emotions to varying degrees. This distribution provides a more comprehensive representation of the utterance's emotional state compared to the ground-truth label. It enables models to learn more label-related information, leading to improved performance in dialogue emotion analysis. Our main contributions are summarized as follows:

- We first introduce the label refinement into the ERC task and propose a novel EmoLR method. It dynamically integrates context-sensitive (global information) and speaker-sensitive (local state) information to refine emotion label distribution and supervise model training.
- The refined label distribution (RLD) can improve the ERC performance without requiring external knowledge or modifying the original model structure. Moreover, context- and speaker-sensitive RLD can be applied to both unimodality and multimodality conversation scenarios.
- Extensive experiments demonstrate the benefits of utilizing context-sensitive and speaker-sensitive RLD to improve ERC performance. The label refinement has high generalization which could be used in the existing ERC models. Furthermore, our method outperforms state-of-the-art results on benchmark datasets.

2 Related Work

2.1 Emotion Recognition in Conversations

Emotion detection in conversations has attracted much attention due to the availability of public conversation datasets. Previous research has focused on textual or multimodal data. Hazarika et al. (2018a) and Majumder et al. (2019) used multi-group gated recurrent unit (GRU) (Cho et al., 2014) networks to capture contextual representations of conversations and speaker states. Jiao et al. (2019) constructed the two-tier GRU network for information extraction of tokens and utterances, respectively. Hu et al. (2021) utilized BiLSTM and attention mechanisms to process utterance representations, simulating human cognitive. Ghosal et al. (2019) and Shen et al. (2021) utilized the dependencies among the interlocutors to extract the conversational context. Context propagation in conversations was strengthened by graph neural networks (GNN) (Wu et al., 2020; Tu et al., 2022b). Considering the importance of

external knowledge for understanding dialogue, Zhong et al. (2019) constructed a commonsense-based transformer and selected appropriate knowledge according to context and emotional intensity. Zhu et al. (2021) extracted topic representations with a self-encoder, and obtained dynamic commonsense knowledge from ATOMIC (Sap et al., 2019) through topics. Tu et al. (2023) obtain a latent feature reflecting the impact degree of context and external knowledge on predicted results by contrastive learning. Shen et al. (2021) introduced commonsense knowledge into the graph structure.

In summary, the output (emotion states) of the above works were all computed using one-hot labels for training. In contrast, our approach monitors the model using refined labels, enabling more comprehensive extraction of semantic information regarding emotions in conversation.

2.2 Label Refinement

The process of transforming original logical labels into a label distribution is defined as label refinement (LR) or label enhancement (LE). Müller et al. (2019) proved that LR can yield more compact clustering. Vaswani et al. (2017) used LR in language translation tasks. Song et al. (2020) used LR to regularize RNN language model training by replacing hard output targets. Lukasik et al. (2020) introduced a technique to smooth well-formed correlation sequences for the seq2seq problem. LDL is an effective method for LR. Xu et al. (2021) introduced a Laplace label enhancement algorithm based on graph principles. Zhang et al. (2020) designed a tensor-based multiview label refinement method to obtain more effective label distributions. Additionally, label embedding has shown promise in classification tasks. Zhang et al. (2018) proposed a multi-task label embedding for transforming tasks into a vector-matching problems. Wang et al. (2018) proposed to regard text classification as a label-word joint embedding problem. Bagherinezhad et al. (2018) studied the impact of label attributes and introduced label refinement in image recognition. However, most LE methods require multi-label information, which is not available in ERC datasets. Obtaining true label distributions manually is also challenging. Our proposed method can generate distributed labels in conversations, and LDL can be more widely applied to ERC tasks.

3 Methodology

3.1 Problem Definition

Let M speakers including $Q_1, Q_2, \ldots, Q_M$, and $U_i = [u_{N_1}, u_{N_2}, \ldots, u_{N_x}]$ represent the utterances in the conversation, where N_i is the number of utterances. Each u_i is spoken by the corresponding speaker $Q_j, j \in [1, M]$. The task is to recognize the emotion labels of each u_i using the original emotion label y_i and the refined label RLD_i. To this end, we aim to maximize the following function:

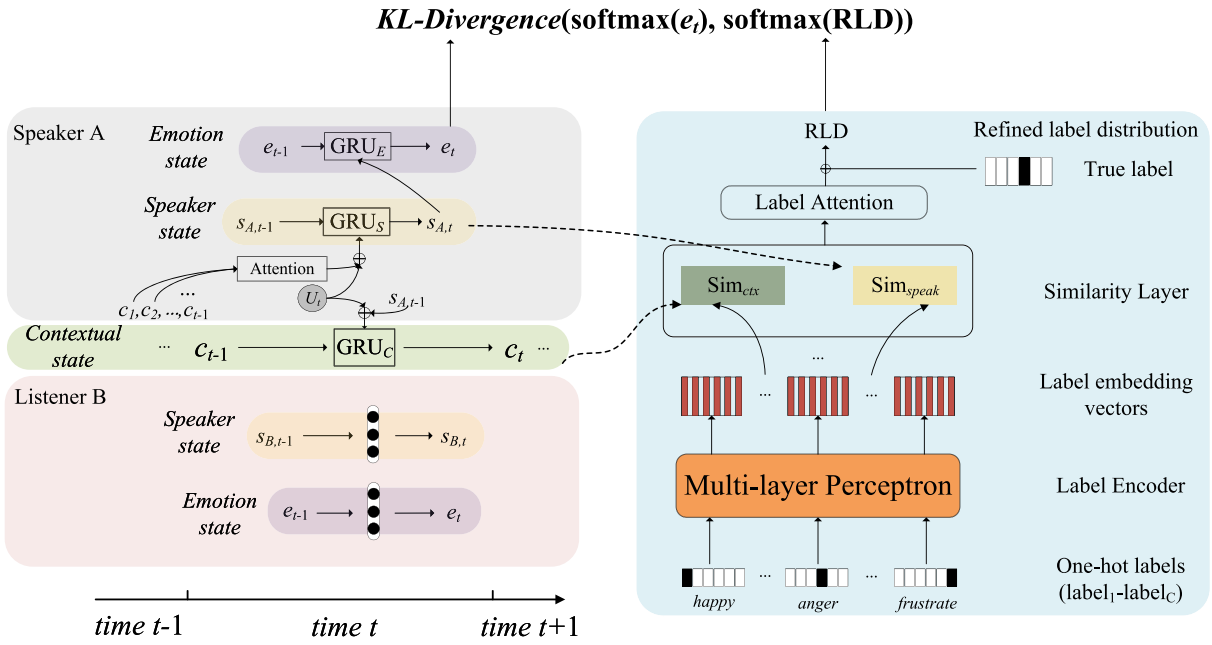
$$f = \prod_{i=1}^{N_x} p(y_i, RLD_i | u_{1..i}; RLD_{1..i-1}; \theta) \quad (1)$$

where RLD_i represents the RLD of the i-th utterance in the conversation, and θ denotes the set of parameter matrices of the model. We propose the EmoLR method to generate the RLD, which provides more label-related information across different emotion classes and improves the performance of the EP.

3.2 Emotion Label Refinement

One-hot labels often ignore important information when utterances convey mixed emotions since they can only represent a single emotion. To learn more label-related information, we aim to obtain a new label distribution that reflects the dependencies among different emotion dimensions within a sample. Considering that context and speaker states are crucial factors for emotions (Hazarika et al., 2018a,b), we propose EmoLR to recover relevant emotional information for training. EmoLR calculates the context-sensitive and speaker-sensitive dependencies between instances and labels respectively, and generates RLD by dynamically fusing two types of dependencies. However, RLD may introduce some noise to the model. Therefore, during training, we integrate both the refined labels and the corresponding one-hot labels to reduce the impact of noise on the model. EmoLR consists of two components: the emotion predictor (EP) and the label refinement (LR). The overall architecture is shown in Figure 2. We will introduce the EmoLR in detail.

The EP is the basic emotion predictor that takes into account both context and speaker states. We aim to simulate the conversation process: The global context information $[c_1, \ldots, c_{t-1}]$ influences the speaker state $s_{p,t}$, and the speaker



(a) Emotion Predictor (b) Label Refinement

Figure 2: Illustration of the proposed model, which is composed of an emotion predictor and a label refinement component. We extract the information from speakers and context within the conversation model to refine the original one-hot label to RLD. Therefore, the label refinement module can be regarded as the main process of EmoLR.

state $s_{p,t-1}$ influences the context state c_t in the next time-step. Finally, the emotion state e_t is updated using speaker state $s_{p,t-1}$ and last time-step emotion state e_{t-1} by GRU network. We utilize three groups of GRU networks, used to extract context state c_t, the speaker states $s_{p,t}$ and the emotion representation e_t, respectively. The process of capturing the representations for the three state is as follows:

$$c_t = GRU_C(c_{t-1}, (u_t \oplus s_{p,t-1})) \quad (2)$$

$$att(\omega_{t-1}) = \frac{\exp(u_i^T W_\alpha \omega_{t-1})}{\sum_{i=1}^{t-1} \exp(u_i^T W_\alpha \omega_{t-1})} \omega_{t-1}^T \quad (3)$$

$$s_{p,t} = GRU_S(s_{p,t-1}, (u_t \oplus att(\omega_{t-1})) \quad (4)$$

$$e_t = GRU_E(e_{t-1}, (u_t \oplus s_{p,t-1})) \quad (5)$$

where u_t represents the t-th utterance in the conversation, ω_{t-1} represents $[c_1, \ldots, c_{t-1}]$, D_u, D_c, D_s, D_e represent the sizes of u_t, c_t, $s_{p,t}$, e_t respectively, D_c, D_s, D_e are set to the same value. $c_t \in \mathbb{R}^{D_c}$ represents the contextual representation at time t, $s_{p,t} \in \mathbb{R}^{D_s}$ represents the state of speaker p at time t, $e_t \in \mathbb{R}^{D_e}$ represents the emotion state at time t, and $W_\alpha \in \mathbb{R}^{D_u \times D_c}$ represents the attention weight matrix.

The LR component consists of a label encoder, a similarity layer, and a label attention module. The multi-layer perceptron (MLP) is used as the encoder to generate the transformation matrix W. The similarity layer takes the label embedding vector, speaker states and context states as inputs and computes their similarity correlation using the dot product. The label attention module is then used to dynamically fuse context-sensitive and speaker-sensitive dependencies to generate the RLD. Thus, the RLD becomes a context-dependent and speaker-dependent distribution that adapts to the choice of context-sensitiveness and speaker-sensitiveness for each emotion. The process can be represented as follows:

$$R^{(l)} = f^e([label_1, label_2, \ldots, label_C]) \quad (6)$$

$$Sim_{speak} = s_p^T R^{(l)} W + b \quad (7)$$

$$Sim_{ctx} = c^T R^{(l)} W + b \quad (8)$$

where f^e is an encoder function that transforms $[label_1, label_2, \ldots, label_C]$ into a label embedding matrix $R^{(l)}$, C is the number of classes,

Dataset	# dialogues			# utterances			# classes	# Metrics
	train	val	test	train	val	test		
IEMOCAP	120	12	31	5810		1,623	6	Weighted-average F1
MELD	1,039	114	280	9,989	1,109	2,610	7	Weighted-average F1
EmoryNLP	659	89	79	7,551	954	984	7	Weighted-average F1

Table 1: The statistics of splits, classes, and evaluation metrics adopted in three different datasets.

W is the transformation matrix, b is the bias, Sim_{speak} is speaker-sensitive score, and Sim_{ctx} is context-sensitive score.

$$H_F = tanh(W_F\,[Sim_{speak}, Sim_{ctx}]) \quad (9)$$

$$\alpha_{att} = w_F^T H_F \quad (10)$$

$$RLD = [Sim_{speak}, Sim_{ctx}]\,\alpha_{att}^T \quad (11)$$

where $W_F \in \mathbb{R}^{D_c \times D_c}$ is the attention weight matrix, $w_F \in W_F$, $\alpha_{att} \in \mathbb{R}^2$ is the attention scores, and RLD represents the refined label distribution based on Sim_{speak} and Sim_{ctx}.

With the LR component, we assume that the learned RLD reflects the similarity relationship between emotions, context, and speaker states. This helps the model more comprehensively represent the emotions in utterances, especially when in the case of mixed emotions.

3.3 Training

During training, the RLD replaces the one-hot label and is considered as the new training target. The model is trained under the supervision of RLD.

To measure the difference between the RLD and the predicted label vector e_t distribution, we use KL (Kullback and Leibler, 1951) divergence as the loss function:

$$y^* = softmax(e_t) \quad (12)$$

$$z = softmax(RLD + y) \quad (13)$$

$$KL(y^*, z) = \sum_{i=1}^{C} z_i \log(\frac{z_i}{y_i^*}) \quad (14)$$

where z is obtained by the softmax of RLD.

Based on the above process, the complete loss function can be written as:

$$Loss = KL + \eta\,\|\theta\| \quad (15)$$

where η indicates the L2 regularization term and θ represents the set of parameter matrices of the model.

It is worth noting that this process occurs only during the training stage and is ignored during prediction. By utilizing this training method, the influence of noise on the model will be reduced as much as possible.

4 Experimental Setups

4.1 Datasets

We conducted comprehensive experiments on three benchmark datasets: (i) IEMOCAP (Busso et al., 2008), (ii) MELD (Poria et al., 2019a), and (iii) EmoryNLP (Zahiri and Choi, 2018). The statistics are shown in Table 1. All these datasets are multi-modal datasets, including textual, visual, and acoustic information for each utterance. However, in this paper, we focus solely on textual information across all datasets.

IEMOCAP is a two-part conversation completed by ten people in five sessions. Each utterance is labeled with one of six emotions: happy, sad, neutral, angry, excited, and frustrated.

MELD is a multi-party dataset collected from the Friends TV series, an extension of the EmotionLines dataset. It includes over 1400 multi-party conversations and 13000 utterances, with each utterance labeled with one of seven emotion labels: anger, disgust, sadness, joy, surprise, fear, or neutral.

EmoryNLP is another dataset based on the Friend TV series. The label of each utterance belongs to one of seven emotion classes: neutral, sad, mad, scared, powerful, peaceful, and joyful.

For all experimental results from these three datasets, we use the accuracy (Acc.) and weighted-average F1 scores (W-Avg F1) as the evaluation

metric to compare the performance of different models.

4.2 Baselines

We compare the performance of EmoLR with the following baselines:

CNN (Kim, 2014) is trained on the utterance level to predict final emotion labels without context information.

ICON (Hazarika et al., 2018a) is a multimodality emotion detection framework. It discriminates the role of the participants.

KET (Zhong et al., 2019) combines external knowledge through emotional intensity and contextual relevance.

DialogueRNN (Majumder et al., 2019) uses two GRU networks to track the state of each participant throughout the conversation.

DialogueGCN (Ghosal et al., 2019) is used to enhance the dependency among speakers of each utterance.

COSMIC (Ghosal et al., 2020) introduces causal knowledge to enrich the speaker states.

ERMC-DisGCN (Sun et al., 2021) proposes to control the contextual cues and capture speaker-level features.

DialogueCRN (Hu et al., 2021) models the retrieval and reasoning process of cognition by mimicking the thinking process of humans, in order to fully understand the dialogue context.

SKAIG (Li et al., 2021) proposes a method called psychological knowledge-aware interaction graph to consider the influence of the speaker's psychological state on their actions and intentions.

DAG-ERC (Shen et al., 2021) utilizes a directed acyclic graph (DAG) to encode utterances and better model the intrinsic structure within a conversation.

4.3 Hyperparameter Settings

For all the baselines, we conducted experiments according to their original experimental settings on the three datasets, using randomly assigned seeds. For our proposed model, EmoLR, we used Adam (Kingma and Ba, 2014) optimization with a batch size of 16, L2 regularization weight of 3e-4, and a learning rate of 1e-4 throughout the training process. The dropout rate was set to 0.5. We used RoBERTa (Liu et al., 2019) to represent word embeddings and took the average value of the last four layers as input. To optimize EmoLR's performance on multiple datasets, we utilized holdout validation with a validation set to conduct a thorough hyperparameter search.

5 Result Analysis

5.1 Comparison with the Baselines

IEMOCAP: On the IEMOCAP dataset, our proposed EmoLR method achieves the best performance among the baselines shown in Table 2, with an average W-Avg F1 score of 68.12%. It outperforms SKAIG by about 1.2% and most other baseline models by at least about 2%. To explain the performance difference, we need to understand the structural features of these models and the nature of the conversation. The top three models in terms of performance are DialogueCRN, SKAIG, and DAG-ERC, all of which attempt to model speaker-level context. However, EmoLR's use of label refinement to encode class information and provide richer context and speaker states than ground-truth labels in the dialogue is a significant reason for its improved performance.

MELD and EmoryNLP: On the MELD dataset, our proposed model achieves a W-Avg F1 score of 65.16%, slightly lower than COSMIC's 65.21%. For EmoryNLP, our model outperforms all baselines except COSMIC and SKAIG by around 2%–4%. The MELD and EmoryNLP datasets, both from the Friend TV series, pose challenges due to short utterances, rare emotion words and many conversations involve more than 5 participants. Emotion-related words rarely appear in these datasets, making it more challenging to design an ideal model. Our proposed model shows better results by efficiently addressing the issues of short utterance and rare emotion words. To capture the complete context and speaker information, the label refinement method utilizes context information and the current local emotion state to compute the global emotion label. However, COSMIC and SKAIG perform slightly better than

Model	IEMOCAP		MELD		EmoryNLP	
	Acc.	W-Avg F1	Acc.	W-Avg F1	Acc.	W-Avg F1
CNN	48.92	48.18	56.41	55.02	–	–
ICON	63.23	58.54	58.62	54.60	–	–
KET	62.77	59.56	59.76	58.18	36.82	34.39
DialogueRNN	64.40	62.75	59.54	56.39	–	–
COSMIC	65.74	65.28	65.42	**65.21**	39.88	38.11
DialogueGCN	65.25	64.18	61.27	58.10	36.98	34.36
ERMC-DisGCN	–	64.10	–	64.22	–	36.38
DialogueCRN	66.33	66.05	60.73	58.39	–	–
SKAIG	–	66.98	–	65.18	–	38.88
DAG-ERC	–	68.03	–	63.65	–	**39.02**
EmoLR	**68.53**	**68.12**	**66.69**	65.16	**42.78**	38.97
$-\text{Sim}_{speak}$	66.42	66.13	65.23	64.17	40.79	37.98
$-\text{Sim}_{ctx}$	66.66	65.20	65.56	63.72	39.05	37.23
−RLD	63.88	63.51	61.82	60.70	38.54	36.52

Table 2: The experimental results on the three datasets. The W-Avg F1 scores are the weighted-average F1 among five runs under different random seeds. Test scores are chosen at best validation scores. '-' represents the reduction of the following part.

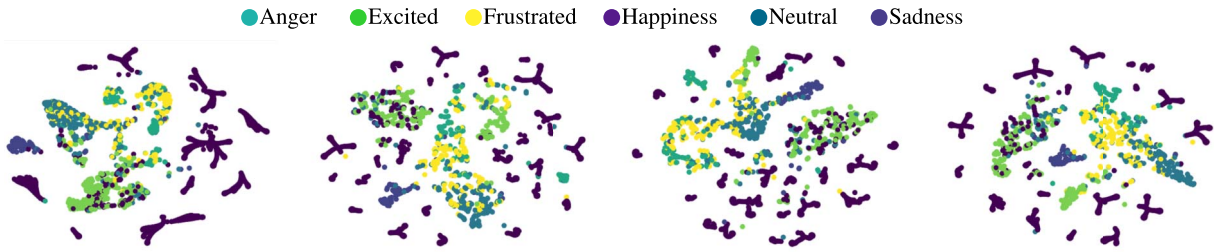
our model by approximately 0.1%, the reason is that they incorporate more specific states with different commonsense knowledge about speakers. Our model only considers the context and current emotion state, making it more suitable for the MELD dataset.

5.2 Ablation Study

In this section, we report on ablation studies to study the impact of different sensitive dependencies in EmoLR, and the results are presented in Table 2. When using only the EP (training without label refinement, EmoLR−RLD), the classification performance is the worst on all three datasets. The W-Avg F1 score is only 63.51%, 60.70%, and 36.52.%, even worse than some baselines. These results highlight the importance of the RLD for the EP. Adding context-sensitive dependence to obtain EmoLR$-\text{Sim}_{speak}$ improves the W-Avg F1 score to 66.13%, 64.17%, and 37.98%, demonstrating the significance of context for label distribution. Similarly, EmoLR$-\text{Sim}_{ctx}$, which represents speaker-sensitive labels without context-sensitive dependence, achieves an F1 score of 65.20%, 63.72%, and 37.23%, proving that speaker-sensitive dependence is essential for label distribution. Notably, EmoLR$-\text{Sim}_{speak}$ is better than EmoLR$-\text{Sim}_{ctx}$, indicating that labels are more sensitive to context. We also observe that some methods yield results closer to our model. To compare them, t-test at a significance level of 0.05 is employed on three comparative methods that showed similar experimental results, as shown in Table 3. On the IEMOCAP dataset, EmoLR significantly outperformed the other comparative methods. On the MELD dataset, EmoLR yielded similar results to COSMIC but surpasses the other comparative methods. On the EmoryNLP dataset, EmoLR performed similarly to DAG-ERC but outperformed the other comparative methods. Overall, EmoLR's performance is superior to most of the comparative methods.

Model	IEMOCAP	MELD	EmoryNLP
EmoLR DAG-ERC	3.793e-2	1.835e-3	4.007e-1
EmoLR SKAIG	1.220e-4	3.856e-3	5.635e-3
EmoLR COSMIC	2.117e-6	7.421e-1	8.438e-3

Table 3: The results of significance test among models with similar performance.



(a) The prediction with r_1 training

(b) The prediction with r_2 refinement

(c) The prediction with r_3 refinement

(d) The prediction with r_4 refinement

Figure 3: t-SNE visualization of the emotion states of utterances from the test sets of IEMOCAP. Colors indicate the ground-truth emotion labels. r_1 represents the one-hot label, r_2 represents the EP without speaker-sensitive refinement, r_3 represents the EP without context-sensitive refinement, r_4 represents the RLD.

Model	IEMOCAP	MELD	EmoryNLP
EmoLR $-\text{Sim}_{speak}$	2.070e-10	3.863e-7	4.137e-6
EmoLR $-\text{Sim}_{ctx}$	1.376e-9	1.677e-7	2.355e-7
EmoLR $-\text{Sim}_{RLD}$	2.914e-13	2.292e-10	1.428e-10

Table 4: The results of significance test in the ablation study.

For clearer presentation and comparison, a visual experiment is conducted to exhibit the predicted distribution with and without label refinement. The t-SNE (Van der Maaten and Hinton, 2008) method is used to demonstrate the visual result by transforming the high-dimension data space into a low-dimension data space. From Figure 3, it is obvious that the distribution of the EP under one-hot label supervision is discrete and chaotic in subfigure (a), making it difficult to find relationships among emotions. In contrast, the prediction distribution of the EP under refined label supervision is significantly closer for the same emotion class. In subfigure (d), the same class exhibits a more compact cluster and a more regular distribution.

Besides, a t-test at a significance level of 0.05 is employed to measure the ablation study. The experimental results are shown in Table 4, and it is obvious that there is a difference between the RLD and the ground-truth label.

5.3 Generalization Analysis

By applying the LR component to other models, such as DialogueRNN, DialogueGCN, COSMIC, and DAG-ERC, we observed significant improvements compared to the baselines presented in Table 5. This demonstrates that LR is not only effective in the EP but also in other models. Based on these experiments, we can conclude that our proposed method is effective and applicable in ERC. Furthermore, Table 6 displays the experimental results from multimodal ERC models. Specifically, we extracted visual and audio information from multimodal data using previous works, such as ICON (Hazarika et al., 2018a) and DialogueRNN (Majumder et al., 2019). Both of these works utilized multimodal information from the same dataset in their experiments. We combined the representations of the three modalities to create a new representation of the corpus, which was then fed into our model. The results show that RLD achieves remarkable performance in both unimodal and multimodal settings, demonstrating its applicability to both text modality and multimodality scenes.

Model	IEMOCAP	MELD
	W-Avg F1	W-Avg F1
DialogueRNN	62.75	56.39
DialogueRNN + RLD	**64.57**	**60.28**
DialogueGCN	64.18	58.10
DialogueGCN + RLD	**65.83**	**61.39**
COSMIC	65.28	65.21
COSMIC + RLD	**67.48**	**65.92**
DAG-ERC	68.03	63.65
DAG-ERC + RLD	**68.76**	**64.01**

Table 5: The results of generalization analysis. After performing a paired t-test ($p < 0.05$), a statistically significant difference was found between RLD and original model.

Modality	IEMOCAP		MELD	
	W-Avg F1		W-Avg F1	
	One-Hot	RLD	One-Hot	RLD
Text	63.88	68.12	61.14	65.16
Audio	52.07	54.90	40.56	42.32
Visual	40.62	42.76	32.77	32.90
Audio + Visual	54.71	58.65	40.29	42.36
Text + Audio	64.79	68.35	62.68	65.20
Text + Visual	64.23	68.20	62.60	65.02
Text + Audio + Visual	65.87	**68.92**	62.83	**65.32**

Table 6: Comparison of the performance on both IEMOCAP and MELD considering different modality combinations. Different modalities was found to be statistically significant under the paired t-test ($p < 0.05$).

Model	IEMOCAP	MELD
	W-Avg F1	W-Avg F1
Emotion Predictor + RLD	**68.12**	**65.16**
Emotion Predictor + LS	65.60	62.87
Emotion Predictor + LCM	66.08	64.10

Table 7: Comparison with label smoothing (LS) and label confusing model (LCM) (Guo et al., 2021).

Model	IEMOCAP	MELD
	PCC	PCC
One-Hot & Human evaluation	0.792	0.708
RLD & Human evaluation	**0.897**	**0.820**

Table 8: Correlation analysis. PCC represents the Pearson correlation coefficient.

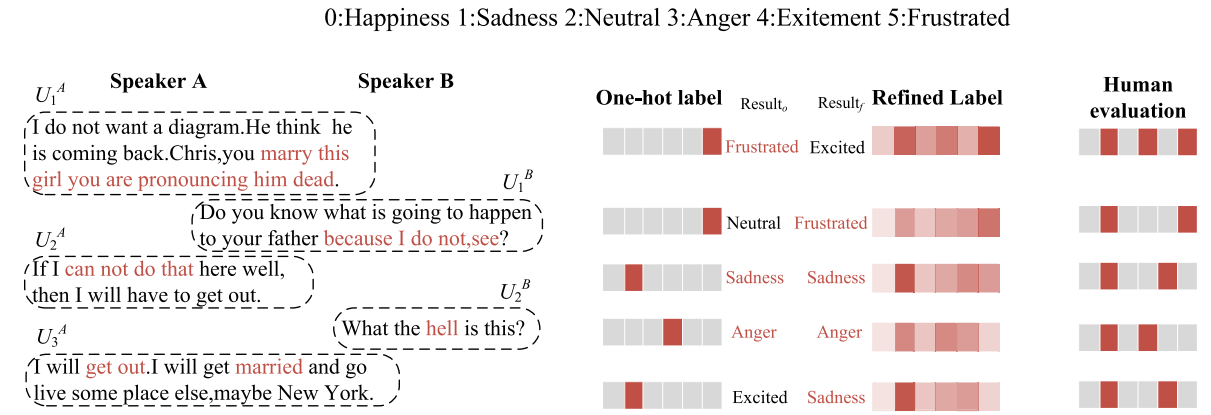
5.4 Comparison with Label Smoothing and Label Confusion Learning

We compared our RLD with other label enhancement methods implemented on the same predictor, and the experimental results are presented in Table 7. It can be observed that RLD outperforms Label Smoothing (LS) and Label Confusing Model (LCM) on both datasets. LS, which randomly adds noise, does not fundamentally solve the weakness of one-hot labels in quantitatively representing corresponding emotions. LCM, on the other hand, is not sensitive to the context and speaker state in the conversation. This constitutes the primary reason for the superior performance of EmoLR.

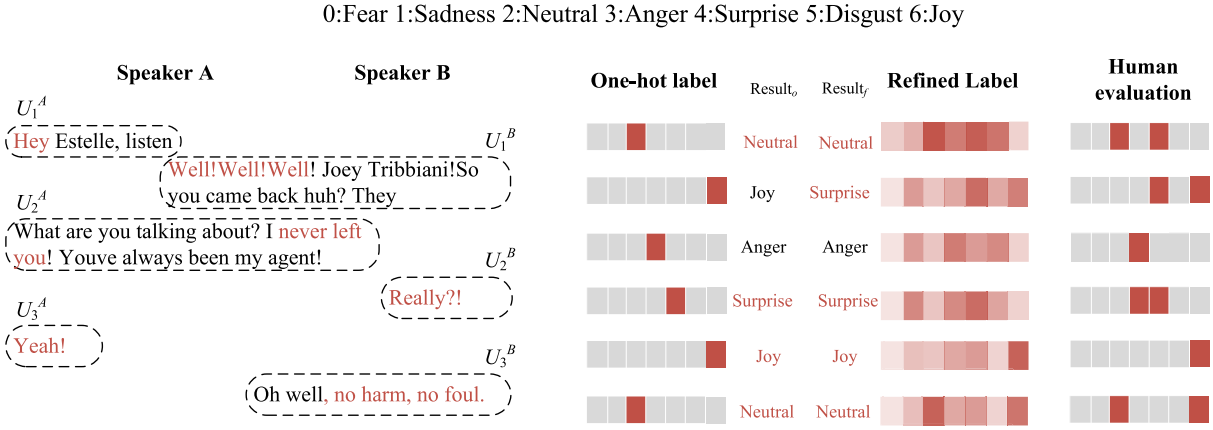
5.5 Correlation Analysis and Case Study

We conducted manual evaluations of each utterance with the assistance of three annotators. During the tagging process, annotators labeled 1 if there was a possible emotion present and 0 otherwise. For example, if an utterance conveyed both happiness and excitement and its label is $\{1,0,0,0,1,0\}$ in IEMOCAP. In cases where there were discrepancies among the annotators, a majority vote was taken, labeling 1 for emotions that received two or more votes, and 0 otherwise. Correlation analyses were performed between the manual evaluations, one-hot labels, and RLD, and the experimental results are shown in Table 8. It can be observed that manual evaluations exhibit a stronger correlation with RLD compared to one-hot labels.

To provide a better analysis of how RLD enhances the emotion detection ability of the EP, we present case studies with selected examples from the IEMOCAP and MELD datasets in Figure 4. Correct results and emotional keywords are highlighted in red. The ground-truth label (Result_o) and refined labels (Result_f) for each utterance are presented separately. In Figure 4(a), the conversation occurs in a negative atmosphere, but U_1^B is a common greeting about the other speaker's father, and the words have no obvious emotional tendency. This contrast with the context leads the model to misjudge and classify the speaker as neutral, assuming that the speaker is not frustrated enough at this point. However, the



(a) A case in IEMOCAP



(b) A case in MELD

Figure 4: Case studies of basic predictor with label refinement.

refined labels consider multiple emotions based on speaker-sensitive and context-sensitive perceptions, enabling the model to detect frustration. Figure 3 also demonstrates that different emotion states trained by RLD overlap less compared to one-hot labels. Similarly, for utterance U_3^A, the high probability of marriage being associated with happiness, combined with a few negative emotional words, leads to an incorrect identification as excited. However, considering the context of sadness, the predominant emotion is correctly associated with sadness, leading to the correct result. In the MELD (b) dataset, the word 'well' is typically associated with a joyful emotion, but in this case, it expresses a more excited state of mind. The RLD results align more closely with human evaluations in both examples, highlighting the effectiveness of label refinement, especially in confusing and partial conditions.

5.6 Error Analysis

Despite the excellent performance of the EmoLR method, there are instances where it fails to recognize certain emotions in dialogues. As shown in Figure 4, our model misclassifies sad utterances as excited. This is because some utterances convey a single but extremely strong emotion, and the refined labels may introduce noise, resulting in incorrect emotion recognition. For example, the frustration emotion in utterance U_1^A is clearly obvious due to the keyword ''dead'', but after label refinement, the result is influenced by the word ''marry'', resulting in an incorrect excited emotion. Figure 3 also shows that many different emotion classes are close in location on the t-SNE result. For example, the clusters of happiness and excitement are located close together, indicating their similarity in original label value.

Label	W-Avg F1
RLD	64.76
RLD + 1*y	65.49
RLD + 2*y	67.12
RLD + 3*y	67.77
RLD + 4*y	**68.12**
RLD + 5*y	67.64

Table 9: An effect of the ground-truth label to RLD on IEMOCAP.

We also explore the impact of ground-truth labels on RLD in Table 9. The experiments show that training the model directly with RLD is not ideal, indicating that RLD is not completely correct. EP still requires ground-truth labels to help mitigate the noise caused by RLD. Notably, the experiment with the ground-truth set to 4 yields the best results, emphasizing the importance of striking the right balance. How to alleviate the negative effect caused by refined labels is a challenge for future work.

6 Conclusion

In this paper, we propose the EmoLR method to adaptively generate a refined label distribution that quantitatively describes emotional intensity. RLD guides the model to learn more label-related knowledge and capture more comprehensive semantic information. Our proposed method influences RLD through context- and speaker-sensitive states without requiring external knowledge or changing the original model structure. EmoLR has been proven effective on both unimodality and multimodality data through extensive experimental analysis and outperforms state-of-the-art results on three datasets.

Future work on EmoLR should address RLD's noise generation with unrelated emotions and reduce interference. Additionally, exploring efficient methods to encode labels in multimodal settings can shed light on mixed emotions' relevance.

Acknowledgments

The authors would like to thank Yang Liu, Cindy Robinson, and other anonymous reviewers for detailed comments on this paper. This research is funded by the National Natural Science Foundation of China (62372283, 62206163), Natural Science Foundation of Guangdong Province (2019A1515010943), The Basic and Applied Basic Research of Colleges and Universities in Guangdong Province (Special Projects in Artificial Intelligence)(2019KZDZX1030), 2020 Li Ka Shing Foundation Cross-Disciplinary Research Grant (2020LKSFG04D), Science and Technology Major Project of Guangdong Province (STKJ2021005, STKJ202209002, STKJ2023076), and the Opening Project of GuangDong Province Key Laboratory of Information Security Technology (2020B1212060078).

References

Hessam Bagherinezhad, Maxwell Horton, Mohammad Rastegari, and Ali Farhadi. 2018. Label refinery: Improving imagenet classification through label progression. *ArXiv*, abs/1805.02641.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359. https://doi.org/10.1007/s10579-008-9076-6

Erik Cambria, Andrew Livingstone, and Amir Hussain. 2012. The hourglass of emotions. *Cognitive Behavioural Systems*, pages 144–157. Springer. https://doi.org/10.1007/978-3-642-34584-5_11

Erik Cambria, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. New avenues in opinion mining and sentiment analysis. *IEEE Intelligent Systems*, 28(2):15–21. https://doi.org/10.1109/MIS.2013.30

Iti Chaturvedi, Ranjan Satapathy, Sandro Cavallari, and Erik Cambria. 2019. Fuzzy commonsense reasoning for multimodal sentiment analysis. *Pattern Recognition Letters*, 125:264–270. https://doi.org/10.1016/j.patrec.2019.04.024

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares,

Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*. https://doi.org/10.3115/v1/D14-1179

Xin Geng. 2016. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748. https://doi.org/10.1109/TKDE.2016.2545658

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481. https://doi.org/10.18653/v1/2020.findings-emnlp.224

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164. https://doi.org/10.18653/v1/D19-1015

Biyang Guo, Songqiao Han, Xiao Han, Hailiang Huang, and Ting Lu. 2021. Label confusion learning to enhance text classification models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12929–12936. https://doi.org/10.1609/aaai.v35i14.17529

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604. https://doi.org/10.18653/v1/D18-1280

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, pages 2122–2132. NIH Public Access. https://doi.org/10.18653/v1/N18-1193, PubMed: 32219222

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7042–7052. https://doi.org/10.18653/v1/2021.acl-long.547

Dazhi Jiang, Hao Liu, Runguo Wei, and Geng Tu. 2023. Csat-ftcn: A fuzzy-oriented model with contextual self-attention network for multimodal emotion recognition. *Cognitive Computation*, pages 1–10. https://doi.org/10.1007/s12559-023-10119-6

Dazhi Jiang, Runguo Wei, Jintao Wen, Geng Tu, and Erik Cambria. 2022. Automl-emo: Automatic knowledge selection using congruent effect for emotion identification in conversations. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2022.3232166

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R. Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *EMNLP*, pages 1746–1751. https://doi.org/10.3115/v1/D14-1181

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Solomon Kullback and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86. https://doi.org/10.1214/aoms/1177729694

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1204–1214.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Michal Lukasik, Himanshu Jain, Aditya Menon, Seungyeon Kim, Srinadh Bhojanapalli, Felix Yu, and Sanjiv Kumar. 2020. Semantic label smoothing for sequence to sequence problems. *Conference on Empirical Methods in Natural Language Processing*, pages 4992–4998. https://doi.org/10.18653/v1/2020.emnlp-main.405

Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32. https://doi.org/10.1609/aaai.v32i1.12048

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11).

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825. https://doi.org/10.1609/aaai.v33i01.33016818

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? *Advances in Neural Information Processing Systems*, 32.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of Emotion*, pages 3–33. Elsevier. https://doi.org/10.1016/B978-0-12-558701-3.50007-7

Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. https://doi.org/10.18653/v1/P19-1050

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035. https://doi.org/10.1609/aaai.v33i01.33013027

Prakhar Saxena, Yin Jou Huang, and Sadao Kurohashi. 2022. Static and dynamic speaker modeling based on graph neural network for emotion recognition in conversation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 247–253. https://doi.org/10.18653/v1/2022.naacl-srw.31

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1551–1560. https://doi.org/10.18653/v1/2021.acl-long.123

Minguang Song, Yunxin Zhao, Shaojun Wang, and Mei Han. 2020. Learning recurrent neural network language models with context-sensitive label smoothing for automatic speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6159–6163. https://doi.org/10.1109/ICASSP40776.2020.9053589

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2949–2958. https://doi.org/10.18653/v1/2021.findings-emnlp.252

Geng Tu, Bin Liang, Dazhi Jiang, and Ruifeng Xu. 2022a. Sentiment- emotion- and context-guided knowledge selection framework for emotion recognition in conversations. *IEEE Transactions on Affective Computing*, pages 1–14. https://doi.org/10.1109/TAFFC.2022.3223517

Geng Tu, Bin Liang, Ruibin Mao, Min Yang, and Ruifeng Xu. 2023. Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14054–14067, Toronto, Canada. Association for Computational Linguistics.

Geng Tu, Jintao Wen, Hao Liu, Sentao Chen, Lin Zheng, and Dazhi Jiang. 2022b. Exploration meets exploitation: Multitask learning for emotion recognition based on discrete and dimensional models. *Knowledge-Based Systems*, 235:107598. https://doi.org/10.1016/j.knosys.2021.107598

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint embedding of words and labels for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Vol 1*, pages 2321–2331. https://doi.org/10.18653/v1/P18-1216

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S. Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24. https://doi.org/10.1109/TNNLS.2020.2978386, PubMed: 32217482

Ning Xu, Yun-Peng Liu, and Xin Geng. 2019. Label enhancement for label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1632–1643. https://doi.org/10.1109/TKDE.2019.2947040

Ning Xu, Yun-Peng Liu, and Xin Geng. 2021. Label enhancement for label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 33:1632–1643. https://doi.org/10.1109/TKDE.2019.2947040

Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Fangwen Zhang, Xiuyi Jia, and Weiwei Li. 2020. Tensor-based multi-view label enhancement for multi-label learning. In *IJCAI*. https://doi.org/10.24963/ijcai.2020/328

Honglun Zhang, Liqiang Xiao, Wenqing Chen, Yongkun Wang, and Yaohui Jin. 2018. Multi-task label embedding for text classification. *Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/D18-1484

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176. https://doi.org/10.18653/v1/D19-1016

Zebin Zhong, Shiqi Yang, and Gary Becigneul. 2021. Environment and speaker related emotion recognition in conversations. In *The 2nd International Conference on Computing and Data Science*, pages 1–6. https://doi.org/10.1145/3448734.3450913

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582. https://doi.org/10.18653/v1/2021.acl-long.125